DOCUMENT RESUME

ED 092 025 HE 005 585

AUTHOR	Kulik, James A.; Ericksen, Stanford C.
TITLE	Evaluation of Teaching. Memo to the Faculty; Memo No. 53 February 1974.
INSTITUTION	Michigan Univ., Ann Arbor. Center for Research on Learning and Teaching.
PUB DATE	Feb 74
NOTE	6p.
AVAILABLE FROM	The Center for Research on Learning and Teaching, The University of Michigan, 109 E. Madison, Ann Arbor, Michigan 48104 ($0.50)
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	*College Faculty; *Evaluation Methods; Evaluation Techniques; Faculty Evaluation; *Higher Education; *Teacher Evaluation; *Teaching Quality
IDENTIFIERS	*University of Michigan

ABSTRACT

Departments, schools, and colleges within the University of Michigan differ widely in the procedures used to evaluate members of their faculties. In some instances student rating forms are used but practices differ as to the kind of information obtained and the weight given to these assessments. Each set of procedures is designed to serve local purposes; this report summarizes research findings that should be considered when interpreting evaluative data already on hand or when mapping out new arrangements for evaluating teachers. Emphasis is placed on student ratings, sources of variation in student ratings, the use of student ratings, ratings by colleagues, ratings by administrators, self-ratings, limitations of performance measures, and student achievement and student ratings. (MJM)

ED 092025

EVALUATION OF TEACHING

The impact on students of outstanding teachers lasts through the years and into settings far removed from the classroom. Over time, however, these influences are difficult to isolate and to measure and so, for practical purposes, the formal evaluation of teachers is focused on the here-and-now. Even so, the task is difficult. Judging the quality of instruction is a far too complicated process to be based solely on an administrator's personal assessment, or on the consensus judgment of peers, or on the filling out of rating forms by students, or on an examination of the course syllabus, tests, and other instructional materials prepared by the teacher. Each of these is useful but none is sufficient. Teaching is an omnibus profession but each teacher is an idiosyncratic person, and many yardsticks are needed to measure competence in this role. Care must be taken to establish the criteria appropriate for each instructional setting and to judge the teacher within this context.

Departments, schools and colleges within this University differ widely in the procedures used to evaluate members of their faculties. In some instances student rating forms are used but practices differ as to the kind of information obtained and the weight given to these assessments. Each set of procedures is designed to serve local purposes and it would be presumptuous for this *Memo* to outline a uniform system of teacher evaluation. Rather, the present report will summarize research findings which should be considered when interpreting evaluative data already on hand or when mapping out new arrangements for evaluating teachers. We are fortunate in having access to a current review of the literature on teacher evaluation by Dr. James A. Kulik, Acting Director, CRLT, and except for the box on pages 5–6, the rest of this *Memo* is a condensation of his manuscript. — SCE

* * *

Professors approach their jobs variously; as scholars, researchers, disciplinarians, healers, managers, and molders. They teach to different audiences, and hold their audiences to different degrees. In the subtle art of teaching, cues as to success and failure take many forms. Students likewise take different approaches to their studies. They learn different amounts in their courses, and rate their teachers with varying degrees of tolerance. Variety is the substance, not the spice, of college life.

The research question in the evaluation of teachers concerns the relationship among t[illegible]e varying but measurable quantities. How much do actual diffe[illegible]ces among teachers contribute to differences in the ratings that they receive? How much do teacher differences influence student learning? Which are the teacher differences that make the difference?

STUDENT RATINGS

In evaluating teachers, it is commonplace to rely on judgments made by observers: a teacher's students, colleagues, superior, or the

Memo to the Faculty

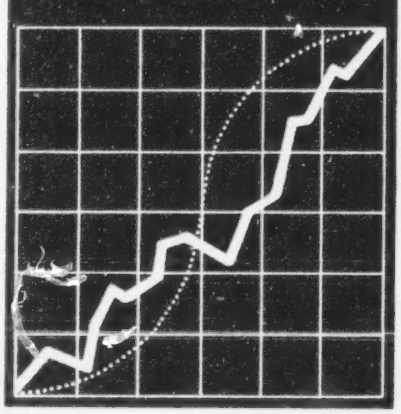
From the Center for Research on Learning and Teaching

The University of Michigan

NO. 53
FEBRUARY 1974

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
University of Michigan
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.



HE 005 585

teacher himself. Student ratings are easiest to obtain, and students have obvious assets as judges of teaching. They are the audience for whom the teaching is intended; they see the teacher day in and day out and are in a unique position to judge a teacher's effectiveness. Many commentators feel that this daily exposure to the teacher's performance more than makes up for the student's lack of age and experience.

Numerous investigators have asked how students judge their teachers, and there is general agreement about the answer. Students see their teachers as differing along several dimensions: (1) in overall teaching skill; (2) in the degree of rapport they establish in the classroom; (3) in the organization and structure of their courses; and (4) in the amount of work they require. While these four factors (Skill, Rapport, Structure, and Overload) are commonly found, the Skill dimension is without question the overriding quality to which student judges react when making an evaluation. "Skill" scores correlate positively and strongly with such general ratings as "all around teaching ability" or "all around value of the course."

Table I

Summary of Determinants of Variation in Student Ratings of Instruction

I. Student variables: The student's general disposition toward instructors and instruction is the most important influence on within-class differences in ratings. Sex, age, grades, and major are of trivial importance.

II. Teaching conditions: Variables which influence class ratings are class size, elected vs. required status of course, and discipline or department of course. While subject matter differences in class ratings within departments have not been demonstrated, this is a likely further source of variation in class ratings, e.g., the teacher of the modern novel may enjoy an advantage over the medievalist.

III. Teacher characteristics: There is probably a weak, positive correlation between experience or academic rank and student ratings, although the size and direction of this relationship may differ somewhat at different types of schools. Research productivity of faculty members shows a similar weak, positive relationship to student ratings. Also highly-rated instructors strike students (as well as peers) as generally cultured and sophisticated and especially as being articulate as classroom lecturers.

IV. Interaction effects: Morsh and Wilder (1954) conclude that if the instructor teaches for the bright students, he will be approved by them and there will be a positive correlation between ratings and grades; if he teaches for the weaker students, he will be disapproved by the bright students and a negative coefficient will be obtained. There is some evidence that college students with different personality traits respond differently to highly structured and less structured teaching styles.

Sources of Variation in Student Ratings

It may seem self-evident that a good teacher will get good overall ratings from his students, and a poor teacher will get poor ratings. But much more than good teaching goes into a good rating. The research findings summarized in Table 1 show that a number of factors influence a student's view of his teacher. Some students, for example, are more severe critics than others and their ratings reflect their overall disposition. It is also clear that the conditions under which an instructor teaches and his/her own personal qualities make a difference in the ratings given by students.

There is some consistency in results of studies relating teacher characteristics to student ratings. Relationships are small or nonexistent between rated teaching effectiveness and general measures of a teacher's knowledge, ability, research productivity, or scholarly traits. The teacher who is rated effective by students differs from the noneffective teacher, however, on measures of communication ability. The highly-rated teacher is verbally fluent and strikes his peers as cultured and sophisticated. He is expressive and enthusiastic. Items describing such an effective communicator are prominent in the evaluative scales of every student rating form. Such items are so salient that evaluative scales—while usually labeled Competence, Skill, and so on—have been labeled Communication or Teacher's Presentation by some investigators. The good teacher is a good talker.

The prevailing forms in higher education today—lectures and discussions—demand verbally-effective educators. The student role in these forms of education is largely that of audience. The verbally-fluent educator who can hold an audience for an hour or two each day is naturally rated more effective than his less verbal colleagues. If the forms of education change in the future and the role of oral presentations becomes less salient, the correlates of rated teaching effectiveness may change.

The Use of Student Ratings

The first investigators to study ratings of teaching concluded that the reliability of well-constructed rating forms was adequate. This conclusion has weathered the years like the Rock of Gibraltar, and it is now generally agreed that the responses of individual students to commonly-used rating forms are both internally-consistent and fairly stable over time. How, then, are these instruments used?

Student ratings can serve various functions: providing administrators

with information for their use in promotion and hiring; giving feedback to teachers for their use in course revision; providing information to students for their use in course selection; and sensitizing all involved to the processes and purposes of teaching and learning.

There are undoubtedly many factors that contributed to administrative caution in the use of student ratings in hiring and promotion. One of these factors can hardly escape the researcher's attention, and that is the non-teacher factors that contribute to variance in class ratings. If factors such as class size, discipline, and course content affect course ratings—as they appear to—then student ratings reflect more than teacher skills. Unless such factors can be taken into account, the use of student ratings for hiring and promotion can be misleading.

Administrators can bypass this difficulty by comparing an individual teacher's ratings with locally-developed norms. In the case of large courses consisting of many sections, one per teacher, such norms can be developed easily. It is more difficult to develop adequate norms for evaluating the hundreds of instructors who teach in non-sectioned courses. To evaluate such teachers, comprehensive, college-wide data on courses and students would have to be collected and used as the basis for norms for courses classified according to size, department, content, elected or required, etc.

It is also possible for individual teachers to use their ratings to improve their courses, but there is little evidence that such ratings are effective aids for the improvement of teaching. Centra's (1972) results are typical. His study on student feedback involved five different types of colleges. On each of 23 items of a rating questionnaire, end-of-semester ratings of teachers who received earlier feedback were nearly identical to those of teachers not receiving feedback. Likewise, instructors who received midsemester feedback did not modify their teaching practices before the end of the course and these results were consistent for instructors in all disciplines, from both sexes, and with varying amounts of teaching experience. In spite of these negative results, Centra concluded that student ratings have some value in improving instruction. He based this view on his finding that at least those teachers who had unrealistically high opinions of their teaching practices at midsemester were affected by feedback, improved their teaching practices, and received improved ratings at the end of the semester.

RATINGS BY PEERS, ADMINISTRATORS, AND ONESELF

A teacher's students are the obvious group to criticize his teaching. Unlike department chairmen, colleagues, and deans, the students have heard and seen the teacher's classroom performances. When colleges evaluate their faculty, however, they consult a teacher's peers and superiors more often than his students. In Astin and Lee's (1967) survey, 85% of all schools used chairmen's evaluations in assessing teacher performance, just over 82% used deans' evaluations, and 49% included colleagues' opinions in such assessments. In contrast, student opinion was utilized in 41% of the schools and systematic student ratings were solicited in only 12% of all institutions. In colleges, but not in courts, hearsay evidence is preferred to testimony from eye-witnesses.

Ratings by Colleagues

Ratings by faculty colleagues agree fairly well with students' ratings of a teacher. Several investigators have looked into this point, and their findings are remarkably similar. This agreement may indicate that teaching skill is a generalized ability. If it is, then a teacher's colleagues could infer his effectiveness at the podium from his performance in faculty-lounge discussions, in debates at department meetings, and in quadrangle conversations. Investigators at the University of Michigan have in fact shown that teaching fellows who are rated highly by students are also considered to be generally artistically sensitive, intellectual, refined, and imaginative by fellow graduate students. To some extent, teaching skill must be a generalized ability.

It should be noted that in this discussion, I have considered ratings that are systematically collected from a teacher's colleagues. Ratings collected and analyzed in this way agree acceptably well with student ratings.

Ratings by Administrators

Insofar as it is possible to judge from published research, ratings of a teacher by college administrators are virtually interchangeable with ratings by the teacher's colleagues. Blackburn (University of Michigan's Center for the Study of Higher Education) and Clark (1971) provide the best data on this point in their study of a small midwestern college. They found a correlation of .62 between ratings by administrators and by colleagues, and this is probably about as high a correlation as the reliabilities of the two composites will allow. These investigators also found a correlation of .47 between administrative

STUDENT RATINGS IN LS&A

In the early 60's, the College asked each teacher to obtain course evaluations from students every third semester. Later, the evaluation procedures were turned over to the students but this proved unsatisfactory. Each department is now expected to carry out its own evaluation program. The Executive Committee of the College insists that the results of student evaluations be included in every department recommendation for promotion to associate or full professor. The college is currently reviewing alternative procedures relating to student evaluation of teachers and courses.

ratings and student ratings of teacher effectiveness. While this correlation is somewhat lower than the correlation between ratings by students and by colleagues, the discrepancy may simply be a function of the different reliabilities of the mechanisms and procedures used to obtain the evaluative data. It seems likely that the correlation between ratings of administrators and of students—if adjusted for unreliability—would be about the same as that between ratings by students and by colleagues.

Self-Ratings

Blackburn and Clark (1971) found little support for the usefulness of self-ratings. They report negligible correlations between self-evaluations and evaluations by administrators or students. The correlation of self-evaluations with ratings by colleagues was also very small. If Blackburn and Clark's result is representative, self-evaluations—whatever they reveal—do not reflect what is commonly taken to be teaching effectiveness.

The problems with self-evaluation are partially a matter of numbers. Ratings by students, colleagues, and administrators are composite ratings. In such composites idiosyncrasies of viewpoint cancel each other out. The severe judgment by one student may be balanced by a more lenient rating from another and, thus, a tempered judgment emerges from disparate views. Self-ratings, on the other hand, are not composites but ratings from a single individual. If the self-evaluating teacher is dull but generous, a poor teacher will get a favorable review; if he is brilliant and self-effacing, a good teacher will get a poor rating.

ACHIEVEMENT MEASURES IN TEACHER EVALUATION

In the sixties, the tide of behaviorism rose once again in the social sciences, swept over educational circles, and flooded journals and textbooks with behavioral terms and concepts. "Behavior management," "behavioral objectives," "behavior modification," and "behavioral measures" became watchwords in education while less behavioral concepts started to drift to sea, to become flotsam and jetsam in an S-R (stimulus-response) world.

Even the area of teacher evaluation was influenced by the behaviorist temper. Rating methods which depend on inferences or judgments by observers came under strong behaviorist attack. McNeil and Popham (1973), for example, concluded their review of teaching competence with the warning that ratings by students, peers, or administrators lack all the desirable attributes of good measures of teaching ability. According to these authors, ratings are contaminated by inference; they do not assess learner growth; and they are collected in non-uniform teaching situations.

Limitations of Performance Measures

Practical difficulties surround the use of performance measures in teacher evaluation measures, and empirical results obtained so far do not encourage high hope. Unless extensive norms have been developed, only instructors giving the same course can ordinarily be compared, and then only when these instructors agree about the content to be covered in their courses. If teachers do not agree on such matters, then any test of student achievement is bound to be somewhat unfair, and differences in class achievement will reflect the degree of unfairness for various teachers.

Even where different sections of the same course are being compared, there is a problem with using student achievement as a measure of teacher effectiveness, and that is the relatively uniform achievement levels of sections taught by different teachers. When students are taught the same subject matter by different teachers, course examinations show that the students often learn fairly similar amounts from these teachers. The point is illustrated in Bendig's (1953) study of student ratings and student achievement. From each student in the introductory psychology classes of five instructors, Bendig obtained an achievement score based on three common examinations and the student's rating of the course and instructor. Bendig found substantial differences between classes in ratings of the course and instructor, but differences between classes in student achievement were trivial and insignificant.

In their reanalysis of data from nearly 100 comparative studies of college teaching, Dubin and Taveggia (1968) document the lack of such achievement differences. These investigators examined reported differences in final examination scores for course-sections taught by different methods, and concluded that the "data demonstrate clearly and unequivocally that there is no measurable difference among truly distinctive methods of college instruction when evaluated by student perfromance on final examinations" (p. 35). At least as measured by performance on examinations, differently-taught sections of courses produce a common outcome.

Hilgard (in Dubin and Taveggia, 1968) has proposed a plausible explanation for the failure to find differences in achievement measures for students taught by different methods. He states:

> Most studies have relied very heavily on a common textbook in all the courses, and, in order to be "fair" most of the examination questions are based on that book. I can't help but believe that more careful exposition goes into a good textbook than a lecturer can put into a lecture. . . . Hence I believe we are often measuring what the student learned from his textbook, which makes it quite indifferent what amusing stories the lecturer told, or how skilled the student was in winning a point in a class argument. (p. 47).

If it is true that textbooks are powerful enough to override differences in teaching methods, it is also very likely true that the textbook will overshadow differences among teachers in flair, skill, and style when the criterion of a teacher's success is his students' achievement.

Student Achievement and Student Ratings

For some time, investigators have tried to relate achievement levels in sections of a course to the average student ratings of the sections. Since it is hard to find real differences among sections in achievement, one would not expect these small differences to relate highly to other characteristics of the sections or instructors.

A recent study by two investigators at the University of California at San Diego (Rodin and Rodin, 1972), however, found a substantial negative correlation (−.75) between student ratings of an instructor and the mean examination score of his students. The report attracted a good deal of attention, and its message was unsettling to many. In the course under consideration at least, students learned least from the teachers they rated most highly.

Several points should be made about this study. First, the N in this study is especially small, and the sampling error of the correlation coefficient is quite large. Second, to a large extent, teaching was the same for all students in all sections of the course. That is, all students met for three days a week for a lecture by the professor in charge of the course, and for all students "the course content was defined by 40 paradigm problems" (p. 1165). Since the students were exposed to the same material and methods for the most part, large differences on examination performance would not be expected. Third, the role played by the unusual criterion measure of student achievement—number of types of problems mastered by students one by one—is essentially unknown. Fourth, no data are presented on section withdrawals and transfers although such factors could easily affect results.

A most important point is that the results of this study are not consistent with those of other investigators. Frey (1973), for instance, investigated the same problem but used a more carefull, controlled design. His correlations between ratings of a teacher's classroom presentations and his exam performance were .91 and .60 for the two calculus courses. Other investigators have reported moderate positive correlations, no correlation, and inconsistent correlations between student achievement and ratings.

In spite of recent enthusiasm about evaluating teachers through their students' achievement, there is little data to suggest that performance measures are really useful in this role. Comparison of achievement measures is practical only in multiply-sectioned courses, and it is hard to see how performance measures can ever provide a common yardstick for ranking faculty members in different fields. A further problem is that students in different sections of a multiply-sectioned course apparently learn fairly similar amounts. Perhaps the most impressive thing about studies relating class achievement to class ratings of instructors is the inconsistency of the results. JAK

A more complete review and bibliography by Dr. Kulik and Dr. Wilbert J. McKeachie will be published in: Kerlinger, F. N. (Ed.), *Review of Research in Education*, Vol. 3. Preprints are available on request from CRLT.

Evaluation and the Individual Teacher

The research studies reviewed by Dr. Kulik involve quantified measures derived from ratings made by groups of teachers, students, and administrators. When viewing oneself in the context of the research analysis of *average* scores or correlations between two sets of *grouped* data, the individual teacher can nearly always take exception—"My case is somewhat different"—and it is.

Teaching is an extremely ego-involved activity and teachers don't like to be categorized and compared any more than do their students. It is interesting to note that a few years ago when the LS&A faculty was considering a formal system of student course evaluation they insisted that the procedure stress qualitative analyses rather than providing a single numerical index of teaching competence—resisting a "grade" is a point well taken whether made by students or teachers.

A teacher's influence on students derives from many factors but in different combination from one teacher to the next or from one course to another. The clarity and relevance of course objectives, for example, should carry considerable weight in teacher evaluation, as should the ability to organize course content into a productive hierarchy, to evaluate students in a manner that supports rather than hinders learning, to provide instructional materials relevant to the objectives of the course, to tutor, to counsel, to arouse students and, finally, to serve as an exemplar or model for the attitudes and values germane to a particular area of research, teaching and public service.

Student ratings can be helpful but contribute only one segment of the evaluation profile. Similarly, the information derived from a classroom visit by a supervisor relates to only part of what it takes to be an effective teacher. The video tape review of teaching performance often limits attention to stylistic manners and maneuvers as the

teacher engages the blackboard, the students, his notes, the scene out the window, the ceiling, and the like. In other words, no single evaluating procedure will do justice to the individuality of the teacher and his long-term impact on students.

Evaluative judgments reflect the values of those making the rating. Deans, department chairmen, and other supervisors view the educational scene from a particular vantage point and their evaluations should emphasize those aspects of teaching consistent with administrative matters—cost effectiveness, housekeeping duties, curricular balance, consistency with institutional goals, etc. One's peers, on the other hand, are in a better position than administrators or students to evaluate such factors as course content, tests and examining procedures, a course outline or syllabus, reference materials and the like. As Kulik points out, students are in a particularly advantageous position to judge the classroom style of the teacher, the interactive environment within a class and, of course, to express their own evaluation of course content.

Student evaluations must be tempered in the light of three rather frequent sources of error: 1) the distortion inherent in the selection of aspects-to-be-rated and the words and phrases used in the item statements. Ideally, each teacher should prepare the rating form to be sure that responses will be received to questions, issues, and about procedures that have personal concern. (Purdue University is developing a procedure which involves a uniform set of evaluating questions to which the individual teacher can add four optional items of his own selection.) 2) comparison against inadequate norms, e.g., a large introductory course has a different instructional climate than the small advanced offering. 3) the fact that unless nearly all students in the course complete the questionnaire (if asked to), the instructor will likely receive a biased sample of opinions.

The less experienced teacher may be disturbed by the fact that despite his best efforts and despite the high ratings he receives from a significant number of students, others in the class see things differently and he is downgraded on several items. These "contrary opinions" have informational value for the teacher. He learns that he cannot be an across-the-board successful teacher; that students will inevitably interact differently with the teacher and with the subject matter of the course.

Quick and easy and superficial evaluating measures tend to degrade the quality of teaching to the level of the assessment procedure. The classic educational aphorism: "The power to test is the power to control the curriculum" applies, in principle, with equal force to the evaluation of teachers. A questionnaire instrument, for example, will not likely serve the purposes of the administration (for promotion and pay), students (as a guide for course selection), and the teacher (as a basis for improving teaching). These different purposes call for particular evaluative procedures. How to tie them all together into a fair and valid profile of the idiosyncratic teacher is the problem that keeps the matter of teacher evaluation on the agenda and an open issue for debate. **SCE**

The Memo to the Faculty *series is published by the Center for Research on Learning and Teaching at The University of Michigan four times each year on a variable schedule. Off-campus subscription rate, $3.00 minimum for the 8 issues #51-58. Single copies $.50; reduced rates 5 or more copies of a single issue.*

This issue was written by Stanford C. Ericksen and James A. Kulik. Inquiries, suggestions, and requests for permission to reprint should be addressed to Managing Editor, CRLT, 109 E. Madison Street, Ann Arbor, Michigan 48104. ©THE UNIVERSITY OF MICHIGAN *1973.*

Memo No. 53

The Center for Research on Learning and Teaching

THE UNIVERSITY OF MICHIGAN
109 E. MADISON
ANN ARBOR, MICHIGAN 48104